InstructDiffusion: A Generalist Modeling Interface for Vision Tasks

Supplementary Material

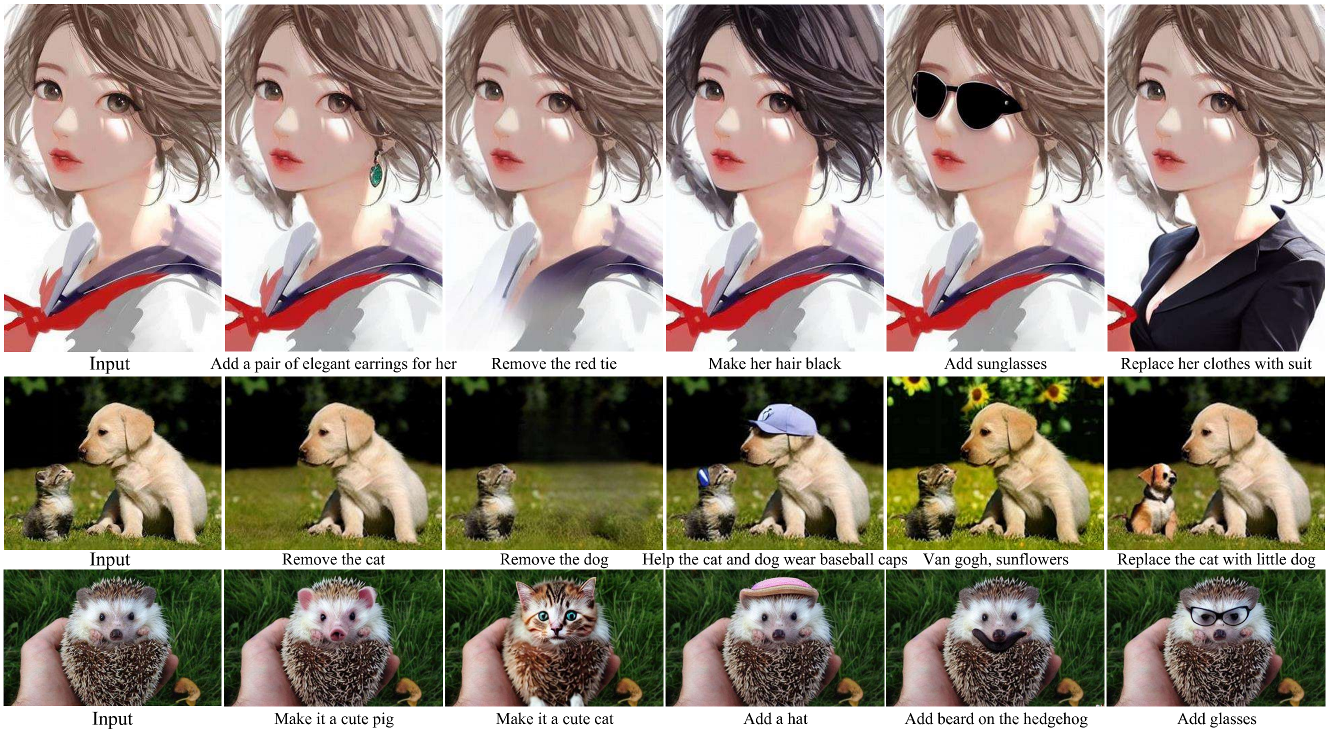

Figure 9. More image editing results generated by our model.

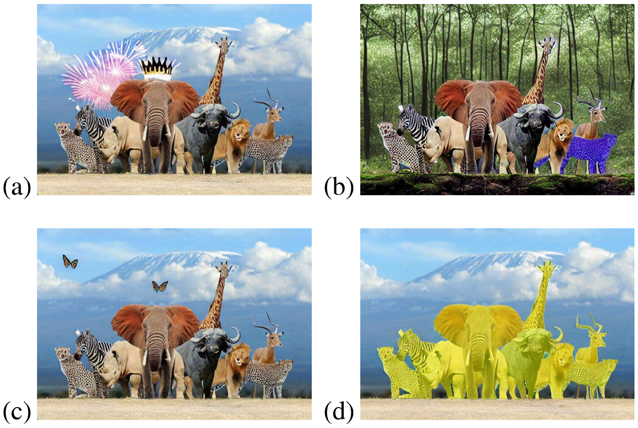

Figure 10. The results of task compositions (a,b) and the failure cases (c,d) generated by our model. The corresponding instructions are as follows: (a) Add **fireworks** to the background and help the **elephant wear a crown**. (b) Change the background into a **forest** and paint the pixels of **right cheetah in blue**. (c) Put **a large butterfly** above the elephant. (d) Mark the pixels of **all the carnivores** to yellow.

A. More visual results for the image editing

Figure 9 illustrates our model's precise editing quality. Our model can add, remove, and replace elements in a source image while maintaining the background integrity and detail preservation.

We further demonstrate the visual results of our model performing task compositions. As illustrated in Figure 10 (a), it is capable of handling multi-task instruction like "Add fireworks to the background and help the elephant wear a crown". For more complex task compositions, we resort to a multi-turn approach, tackling each task one by one, as shown in Figure 10 (b). In future research, we could consider incorporating such task compositions into the training samples. This should potentially enhance the model's capability handling task compositions.

We also present several failure cases of our model in Figure 10 (c-d). Our approach is built upon stable diffusion v1.5, which inherits some of its known drawbacks. First, it may generate images with inaccurate object counts. For example, in Figure 10 (c), two butterflies appear instead of the instructed one. Secondly, it also struggles with complex

Table 7. Ablation study on the pretraining adaptation stage.

Method	COCO	HumanArt	AP-10K
Without pretraining adaptation	71.1	50.9	13.8
With pretraining adaptation	71.2	51.4	15.9

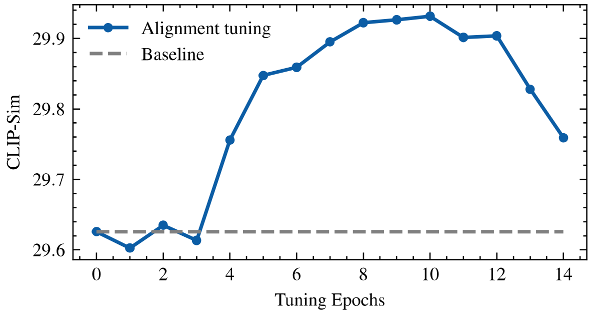

Figure 11. Effect of human alignment. Further fine-tuning on selected human alignment data enhances the CLIP-Sim metric, reaching its peak after approximately 10 epochs.

semantics due to CLIP text encoder constraints, as shown with the term ”carnivores” in Figure 10 (d). Moreover, it loses image quality in high-frequency details due to diffusion in the VQVAE feature space, shown in Figure 10. We believe that as the foundational models like stable diffusion and text encoding improve, our method will see corresponding enhancements.

B. Ablation study on the training stage

Pretraining Adaptation. The output space of Stable Diffusion is comprised of natural images, yet our task-specific training desires the generation of non-natural images with keypoint indicators, diverging from natural images. Therefore, we adopted a pretraining adaptation stage to adjust the output distribution of Stable Diffusion. The results in Table 7 indicate that the pretraining adaptation stage indeed enhances the performance of keypoint detection and proves to be more beneficial for novel datasets.

Human Alignment. To enhance the quality of editing, we conduct the human alignment. Our model, fine-tuned on a curated dataset with human alignment, shows significant image-text alignment improvement, as seen in Figure 11. Initially, the model starts with a CLIP-Sim score of 29.6 and rises to 29.9 over roughly 10 epochs. This notable gain, achieved on a dataset of merely 1,000 samples, underscores the fine-tuning impact on the model’s effectiveness.